Александр Васильевич Зубов (Минск)
Минский государственный лингвистический университет
proscien@mslu.by

ПОДХОДЫ К АВТОМАТИЧЕСКОМУ ИЗВЛЕЧЕНИЮ ТЕРМИНОВ ИЗ ТЕКСТА

Интенсивное развитие современных информационных технологий привело к значительным изменениям в процессах обработки научно-технических и публицистических текстов, процессов обучения и извлечения знаний. И все эти задачи неразрывно связаны с понятием «термин». До сих пор не выработано единое и общепринятое определение понятия «термин». Ориентируясь на автоматическое извлечение терминов из текста, будем считать, что термин – это языковой знак, который может быть и отдельным словом и словосочетанием. Он является носителем элементарной научной, технической, производственной и тому подобной информации в виде отдельного научного понятия, входящего в систему понятий определенной области знания или деятельности. Не выделены и конкретные методы выделения из текстов терминов–слов и терминов–словосочетаний. Предлагаются для этого различные критерии: дефинитивный, критерий концептуальной целостности, информационный критерий, логико-интуитивный и статистический. Для извлечения однословных терминов выделяют два метода: семантический и статистический. Семантический метод выполним только человеком. Он должен иметь для этого специальные словари сочетаемости слов, тезаурусы, энциклопедии и т.п. Статистический метод выделения из текстов терминов может быть реализован автоматически, с использованием ком-

пьютера. Для выделения терминов из текста какого-либо подъязыка, необходимо иметь еще несколько текстов, относящихся к другим подъязыкам или к текстам научно-технического или общественно-политического стиля. В докладе детально рассматриваются два статистических метода автоматического извлечения терминов из текста.